



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이학박사 학위논문

유전체 변이의 임상적 영향 평가를
위한 통합 유전체 데이터베이스

An Integrated Genomic Database for
Evaluating Clinical Impact of Personal
Genome Sequences

2016년 2월

서울대학교 대학원
협동과정 생물정보학 전공
박 찬 희

유전체 변이의 임상적 영향 평가를 위한 통합 유전체 데이터베이스

An Integrated Genomic Database for
Evaluating Clinical Impact of Personal
Genome Sequences

지도교수 김 주 한

이 논문을 이학 박사학위논문으로 제출함

2015년 12월

서울대학교 대학원

협동과정 생물정보학 전공

박 찬 희

박찬희의 박사학위논문을 인준함

2015년 12월

위 원 장 _____ (인)

부 위 원 장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

Abstract

An Integrated Genomic Database for Evaluating Clinical Impact of Personal Genome Sequences

Park Chan Hee
Interdisciplinary Graduate Program in Bioinformatics
Seoul National University

Searching biological databases for interpreting personal genome sequence is the essential routine that interpret the results of biological experiments and form a new hypothesis in genomics and proteomics fields. It is difficult to retrieve all related information despite most researchers build in-houses or local databases to share information within groups. One must retrieve numerous resources to collect biological entries.

With increase of biological data and heterogeneous annotation scheme of genes, integration of gene-centric databases is demanding. Identifying identical genes across different gene-centric databases is a central problem in the integration of various biological databases. Traditional methods of identifying identical genes by gene symbol or genomic location may be often problematic because genes were not uniformly annotated leaving numerous genes not annotated and different methods of gene building for the identical genes can often result in different genomic locations.

We designed reliable and verified schemes to identify identical genes across three gene-centric databases (EntrezGene, UniGene, and Ensembl) using cross reference information, gene symbol and genomic location information. Gene-to-Gene cross reference network (GGN)

was constructed using cross reference information. To increase reliability on identity of genes, GGN went through several procedures using topology of the network, producing reliable gene-to-gene cross reference network (RGGN). RGGN was highly consistent with traditional methods using gene symbol and genomic location. Lists of identical genes could be obtained through the processes of RGGN construction and then validation by gene symbol and genomic location. In contrast to gene integration scheme based on factitiously defined gene concepts, these schemes are natural, data-driven, and clear. Conflicts between different gene-centric databases are resolved by the introduction of network topology. We call this scheme as 'Closed Integration'.

Only considered biological databases with cross-reference information, 'Open Integration' approach integrates cross-reference network around biological databases' identifiers, and resolves the counterpart identifiers in a target database from an input identifier describing how they are connected. This is useful for researchers who need to assess information across multiple databases and for integrating massive biological databases.

Using these schemes, integrated gene-centric database 'GRIP(Genome Resource Annotation Pipeline)' was made, which is combined the benefits of open-closed-integration including OOP modeling as a balanced fashion.

GRIP was modeled ten biological objects divided by three categories, basic, complex, and knowledge to retrieve resources efficiently, Agent-GRIP provides a function that allows searching through the open-closed-integration GRIP. GRIP provides keyword-based search and 'biological knowledge-based search', which is enabled users can

define the desired search by combining each object in GRIP. Besides integration, these schemes can be also used for error corrections of biological databases. This is useful for researchers who need to assess information across multiple databases to interpret microarray experiment results, exome seq, rna seq and personal genome sequence analysis.

**keywords : biological database, database integration,
personal genome sequence**

***Student Number* : 2006-30132**

List of Tables & Figures

Table 1. Matched pairs whose gene symbols are equal but genomic coordinates are completely different.	24
Table 2. Categories, Class, Objects and Attirbutes in GRIP System	41
Figure 1. Strongly Connected Components	21
Figure 2. Flowchart of integrating three database resources	21
Figure 3. (upper) homogeneity at least two BSCC with CC. Y-axis means the homogeneity. (lower) the overlapSimilarity at least two BSCC with CC	23
Figure 4. Density plot of genomic coordinate agreement within BSCC	25
Figure 5. Cross-reference topology around POTEb and POTEc	25
Figure 6. BSCC Examples	26
Figure 7. (a) error case, (b) merge by finding secondary BSCC after condensing primary BSCCs.	27
Figure 8. CC,SCC and BSCC example. Blue box indicates Connected Component	28
Figure 9. Singleton nodes need to be merged into another high similar BSCCs.	29
Figure 10. After merging singleton nodes.	29
Figure 11. Outcome screenshots of ID profiling and ID mapping	32

Figure 12. Overview of GRIP System	36
Figure 13. The Categories of GRIP Objects	38
Figure 14. the 'KCNQ1' as input the results found in the results closed GRIP and the Agent GRIP Display by Object	44
Figure 15. the '1LJD' as input the results found in the results open GRIP and the Agent GRIP Display by Object	45
Figure 16. Knowledge Based Search Scheme.	47

Contents

Abstract	1
Introduction	7
Cross Reference Graph-Based Identicalness of Gene augmented by gene symbol equality and genomic coordinate agreement	12
Guided navigation across biological databases with cross reference visualization	29
GRIP(Genome Research Information Pipeline)	33
Discussion	48
References	50
Abstract (Korean).....	58

INTRODUCTION

An integrated biological database should provide consistent perspectives on entities it describes. Although a few conceptually or technically useful integration schemes to overcome syntactic or semantic heterogeneity were suggested (Kohler, *et al.*, 2003; Stein, 2003; Sujansky, 2001), the more fundamental problem in integration is to present a unified view on each of entities by reconciling contradictory perspectives on each of them and their relationships. The integration of biological data from multiple sources, therefore, cannot avoid the step of identifying the identicalness of entities across different data sources.

“Gene” is one of the most important and well known entity in biology, but the concept of “Gene” is not always clear to database designers who should make databases of genes. The notion of “Gene” in three major gene-centric databases, EntrezGene (Maglott, *et al.*, 2007), UniGene (Schuler, *et al.*, 1996; Wheeler, *et al.*, 2003), and Ensembl (Hubbard, 2002) is different one another. In detail, a human genome consists of 46,491 genes in EntrezGene, 123,641 clusters (an entity equivalent to gene in UniGene) in UniGene, and 44,650 in Ensembl and each gene is often differently annotated among databases about its function, name, and genomic coordinate according to each database’s own scheme. For example, even one of the most famous genes, p53 gene has different ranges of genomic coordinates between EntrezGene and UniGene. The most important barrier in the integration of gene-centric databases is lack of absolute identicalness of genes across those databases.

These problems can be approached in two kinds of strategies. The first is to remove the origin of ambiguities on genes by more concretely and clearly redefining the concept of “Gene”. Then database designers of different gene-centric databases will have identical notion on “Gene”, resulting in identical number and annotations of genes among three gene-centric databases by reorganizing the contents of databases based on a redefined concept of “Gene” so as to remove discrepancies on the properties of genes across databases. This approach is ideal, but does not appear feasible because 1) it is extremely hard to reach a consensus on the definition of “Gene” among biologists that is enough finely granular to remove all the ambiguities on “Gene” and 2) furthermore the definition of “Gene” may be evolving as new biological data and knowledge is generated that cannot be explained by traditional concepts on “Gene”(Gerstein, 2007). Even if “Gene” were ideally redefined, it would not be feasible to change the contents of the databases based on the redefined concept of “Gene” considering the quantity of data accumulated until now and lack of automated procedures to apply the new concepts of “Gene”.

The second approach is to present reasonably compromised, but consistent criteria of identicalness of genes among gene-centric databases and apply it to the task of integration without changing the contents of databases. For example, a gene pair each from EntrezGene and UniGene with a gene symbol “p53” has small regions of non-overlapping genomic coordinates, but most biologists consider it identical. That gene pair can be considered identical by them in spite of the difference of genomic coordinates of that gene pair because that difference is trivial to biologists and their functions are mainly identical. If these biologically meaningful criteria on identicalness on

genes can be consistently applied to all pairs of genes from all gene-centric databases, an integrated database can be made in a feasible way. In this way, identicalness with consistency and biologically meaningful criteria is more useful than the absolute identicalness in the integration of gene-centric databases. Principally this approach can be performed either with or without manual curations by experts, but manual curation is not feasible because the amount of data is too large and the criteria of identicalness is hard to be manually applied consistently to all genes. Instead of it, it is necessary that computationally feasible criteria of identicalness that can be applied to obtainable data on genes are clarified.

Genome-wide data resources that can be used for the construction of biologically meaningful criteria of identicalness of genes across major gene-centric databases include official gene symbol, genomic coordinates, and cross-references from one database to another. Official gene symbol is made by HUGO Gene Nomenclature Committee (HGNC) for human genes and the other equivalent organization for the other species. It can be a strong mediator in the integration of gene-centric databases, but some genes in major gene-centric databases do not have gene symbols—only 30,000 human gene symbols were approved by HGNC while EntrezGene has over 40,000 genes in its database and genes with the same gene symbols in different databases often have completely different genomic coordinates, indicating that those genes are not biologically identical. Genomic coordinate can also be a good indicator for the integration of gene-centric databases because every gene should have a distinct genomic coordinate, but different sequence elements can be

incorporated to identical genes according to assembly methods, so genomic coordinates are hard to be matched across databases.

Major gene-centric databases provide cross-references to other major ones. DAVID knowledgebase, an integrated gene-centered database, was made using cross-reference information of integrated databases(Sherman, 2007). DAVID Gene Concept is a cluster of genes and gene-equivalent entities directly linked by cross-references among many heterogeneous databases including EntrezGene, UniProt UniRef100, and PIR NRef100(Sherman, 2007). Although a very useful approach for integration using cross-reference information was suggested in DAVID knowledgebase, cross-reference information without proper preprocessing may produce weird gene clusters because some cross-references have many errors and broken links.

All of three information sources, gene symbol, genomic coordinates, and cross-reference have important elements to determine identicalness of genes, but simple introduction of each of them is insufficient or incorrect to determine biologically meaningful identicalness of genes.

To make a more concrete and sounder methodology of identicalness of genes, 1) each of three information sources should be properly preprocessed, 2) then preprocessed information should be properly combined to determine identicalness of gene pairs, and 3) those identicalness should be validated.

In this research, we developed a method to identify identical gene pairs each from different gene-centric databases using official gene symbol, genomic coordinates, and cross-reference information. These data sources were independently preprocessed considering the origin of flawed information from each source may be different. Official gene symbol and cross-reference information has a lot of false positive

information, so if we can reasonably remove false positive information, more biologically meaningful identicalness information can be obtained.

In the cases of cross-reference information, filtering of false positives was achieved by analyzing the topology of cross-reference network and finding biologically meaningful sub-networks in it. In contrast to it, absolute identicalness of genomic coordinates between two identical genes is hard to be achievable in many cases, so the process of optimal thresholding of it is required. Because the three information sources were independently collected, each of these three schemes of identifying identicalness can be validated using the other schemes. Combining of three preprocessed information may produce more reliable identical gene pairs because these information can be complementary one another. Finally identical gene clusters were compared to DAVID Gene Concepts for the justification of our method.

Cross Reference Graph-Based Identicalness Of Gene Augmented by Gene Symbol Equality and Genomic Coordinate Agreement

Data sources

We used database resources downloaded 15th July, 2009 and UCSC Genome Browser hg18. Each of EntrezGene, UniGene and Ensembl is a genome-wide information source including gene symbol, genomic coordinate and cross-references for human data only. These data sources were downloaded and localized.

Gene symbol equality-based identicalness

Each human gene in EntrezGene, UniGene, and Ensembl a unique gene symbol given by HGNC. We can understand which functional unit a gene constitutes by its gene symbol. EntrezGene, UniGene, and Ensembl has 45,393, 23,502, and 37,992 unique gene symbols from 46,491, 123,641, and 44,650 total human genes. These gene-centric databases have 64,119 unique gene symbols in total, 22,690 are overlapped between EntrezGene and UniGene, 18,401 between UniGene and Ensembl, 20,007 between Ensembl and EntrezGene, and 18,330 gene symbols are used in all three databases in common. Any genes from the same gene-centric database do not have the identical gene symbols. If a gene pair from two different databases has an identical gene symbols, we considered these genes identical at the aspect of gene symbol and otherwise not identical. Although genes belonging to

a gene family can have very similar gene symbols, we considered them as two different genes.

In spite of providing a fundamental clues for determining the identicalness of genes, gene symbol was not always reliable to identify biological identicalness of genes because some erroneous information is provided. For example, two genes each from EntrezGene and Ensembl with “POTEB” as the gene symbol are located in different chromosomes (chromosome 15 and 18, respectively), indicating that those two genes are not identical biologically.

Although not perfect, gene symbols are very powerful indicator for biological identicalness of genes across different gene-centric databases. Identicalness of gene symbols, therefore, was used to identify identicalness of genes as described in the latter section.

Genomic coordinates agreement-based identicalness

Each gene in each of EntrezGene, UniGene, and Ensemble has distinct regions of genomic coordinate within the database. Agreement of genomic coordinate is also a fundamental property that can determine the identicalness of genes because any gene should have a corresponding genomic coordinate in chromosomes. Although ideally the identicalness of genes should be determined by genomic coordinate agreement, diverse assays and data processings produce many different genomic coordinates for the apparent same gene. The proportion of genes where their regions of genomic coordinates are identical to any genes from other gene-centric databases is only xx.x%, most of biologically identical gene pairs from different gene-centric databases are not absolutely identical in genomic coordinates as described above. Although not absolutely identical, these

gene pairs share significant portions of overlapping in genomic coordinates. In this background, it was necessary to measure the degree of identicalness of gene pairs and to use it as a measurement to determine the identicalness of genes. We defined Genomic Coordinate Agreement (GCA) as in the following and investigated the distribution of it over a genome.

$$GCA(\alpha, \beta) = \arg_{\max} \{Corr(\forall \alpha_{TR}, \forall \beta_{TR})\} \in [0, \dots, 1],$$

$$where \alpha_{TR} \in \{e_{TR}, u_{TR}, en_{TR}\}, \beta_{TR} \in \{e_{TR}, u_{TR}, en_{TR}\}$$

where

e_{TR} is the Reference sequence(s) in e,

u_{TR} is the UniGene sequence in u and

en_{TR} is the Ensembl Transcript ID(s) in en.

and

$$Corr = \frac{cN - ab}{\sqrt{ab(N-a)(N-b)}}$$

a is the number of "bits" (e.g., bases in exons) in data object α , record, b is the total number of bits in all overlapping data object β (within the interval), c is the number of bits in both the α and the β , and N is the length of the interval.

The distribution of GCA for any pairs of each gene from two different gene-centric databases were investigated (Fig). Although not perfect, by optimal thresholding of GCA, we can get more biologically meaningful gene pairs using genomic coordinate.

Integrated cross-reference network

Gene cross-reference network was constructed by integrating all cross-reference information from EntrezGene, UniGene and Ensembl.

EntrezGene provides cross-references to the other two databases. So does Ensembl. But UniGene's cross-reference directs only to EntrezGene. The mutual cross-reference information can be represented as a directed graph with nodes representing genes and directed edges cross-references, creating Integrated Cross-reference Network (ICN) with 214,782 nodes (46,491, 123,641, and 44,650 genes from EntrezGene, UniGene and Ensembl, respectively) and xxxxxxxx directed edges (43,942, 24,510 and 57,200 cross-references from EntrezGene, UniGene and Ensembl, respectively).

Although ICN is a network on the identicalness of genes, the identicalness of genes cannot be not straightforwardly represented from ICN. According to the way of the interpretation of ICN, the identicalness can be differently defined and/or induced. We presented three ways of the interpretation. First two genes on a directed edge are identical regardless of the directions of the edges and multiple edges are connected with "AND" operator. Second a directed edge from node a to node b is interpreted as IF a THEN b and multiple edges are connected with "AND" operator. For example, if there is a path starting from a through b to c, this path states that

IF a THEN b
AND
IF b THEN c.

From the above statement, it is logically inferred that

IF a THEN c.

A bidirectional edge is considered as two edges. In this interpretation, the identicalness can be induced for genes, a and b if a statement containing IF a THEN b AND IF b THEN a is inferred from any paths in ICN. Third a bidirectional edge is a unit of identicalness and multiple bidirectional edges are connected with “AND” operator. A bidirectional edge can be considered as a mutual approval of the identicalness on two genes, so it should be a unit of the identicalness. It is also logical that the identicalness is inferred by the combinations of multiple identicalness with “AND” operator.

All the paths in ICN can be interpreted in these three ways. We will discuss graphical features of identical genes according to each of three ways of interpretations on ICN in the latter section.

ICN-based identicalnesses

From the above-mentioned three ways of interpretation on ICN, three types of identicalness on genes can be defined and/or induced and they show distinct graphical features.

In the first interpretation, a set of identical genes in ICN constitute a connected component (CC). CC is a maximal subgraph where any node is connected to any other nodes regardless the directions of edges. Because directions of the edges were ignored in the first interpretation, it is quite intuitive that a set of identical genes constitutes a CC.

In the second interpretation, a set of identical genes constitutes a strongly connected component (SCC). The identicalness of genes can be demonstrated only when there is a set of cyclic paths containing all of those genes and not containing any other genes. Then any genes in this set can reach any other genes in this set and this set constitutes

a SCC. On the contrary, if a set of genes in ICN constitutes a SCC, all the genes in SCC are identical. In this way, we can find out a set of identical genes from ICN.

A directed graph is strongly connected if there is a path from each node in the graph to every other node. A SCC of a directed graph is a subgraph where all nodes in the subgraph are reachable by all other nodes in the subgraph. Reacheability between nodes is established by the existence of a path between the nodes. If each strongly connected component is contracted to a single node, given a directed graph G , the resulting graph is a directed acyclic graph, the condensation of G .

In the third interpretation, a set of identical genes in ICN constitutes a distinct structure which we would like to define as a bidirectionally strongly connected component (BSCC). For the definition of BSCC, we first define bidirectional path as a sequence of nodes such that from each of its nodes there is a bidirectional edge to the next vertex in the sequence. A directed graph is bidirectionally strongly connected if there is a bidirectional path from each node in the graph to every other node. A BSCC of a directed graph is a subgraph where all nodes in the subgraph are bidirectionally reachable by all other nodes in the subgraph. Bidirectional reachability between nodes is established by the existence of a bidirectional path between the nodes. Bidirectional reachability imposes further constraint on SCC by refusing circular reachability. In Fig. 1, $G(c, d, h)$ satisfies bidirectional reachability for all nodes but $G(a, b, c)$ dose not.

The identicalness of genes can be demonstrated only when all of those genes are connected with bidirectional paths. These genes constitute a BSCC according to the definition of BSCC. On the

contrary, if a set of genes constitutes a BSCC, all the genes are identical.

A directed graph can be decomposed into SCCs by running the depth-first search (DFS) algorithm twice: first, on the graph itself and next on the transpose of the graph in decreasing order of the finishing times of the first DFS. Given a directed graph G , the transpose GT is the graph G with all the edge directions reversed.

A SCC can be decomposed into BSCCs by examining the presence of bidirectional paths.

A BSCC is a subgraph of SCC and SCC is that of CC. From CC to SCC to BSCC, progressively stricter criteria of identicalness is applied. Which criteria of the identicalness of genes is biologically meaningful among CC, SCC, and BSCC was tested using gene symbol and genomic coordinate of genes in the latter section.

Validation of ICN-based identicalnesses by identicalness based on gene symbol equality and genomic coordinate agreement

In the validation of ICN-based identicalnesses, the identicalness based on gene symbol equality and genomic coordinate agreement can be considered as gold standards because they were founded on the more fundamental features of genes. Through this validation, we investigated whether three types of ICN-based identicalness were consistent enough with the identicalness based on gene symbol equality and GCA and which type is the most consistent.

There are three types of ICN-based identicalnesses, CC, SCC, and BSCC. For each of them, we investigated the consistency with the gene symbol equality-based method. We defined consistency between two measurements as in the following.

Consistency = # of shared identical gene pairs between two methods / # of unique identical gene pairs produced by either ICN or gene symbol equality method

If any of consistencies were enough high, we considered those types of ICN-based identicalnesses were validated and one of ICN-based identicalnesses having the highest consistency was the best measurement for the determination of identicalness among the three ICN-based identicalnesses.

In the GCA-based identicalness, the minimal GCA as determined to be 0.9 based on the distribution of GCA over the genome (Fig). We also applied the above-mentioned consistency with GCA-based identical gene sets as gold standard and investigated if they were highly consistent and which type of ICN-based identicalness was the best.

For more validation of ICN-based identicalness using GCA, we investigated the distribution of the GCA of identical gene pairs for each of CC, SCC, and BSCC. If a method is a good measurement on the identicalness, most of gene pairs would have a GCA close to 1. Based on this investigate, we could determine whether any of ICN-based identicalness were good and which type of ICN-based identicalness was best.

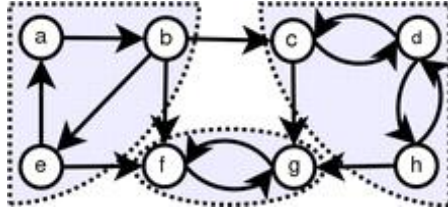


Fig. 1. Strongly Connected Components

Validated identical genes by mutual validation

From the validation test described in 2.6, we found out ICN-based identicalness was validated and BSCC was the best measurement among three types of ICN-based identicalnesses. Then we tried to produce sets of identical genes among three gene-centric databases.

Although all of three measures (BSCC, gene symbol equality, GCA) to produce sets of identical genes were validated, all have low error rates. To make these error rates lower, sets of identical genes produced by one of the three measures were mutually validated by the other two measures and these sets of gene were called validated identical genes.

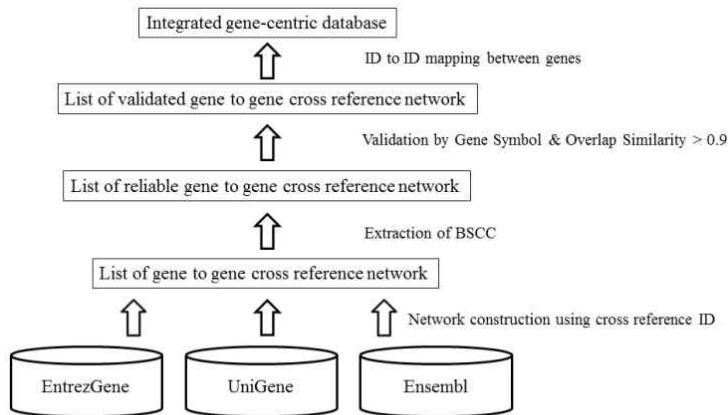


Figure 2. Flowchart of integrating three database resources

Identical genes established by gene symbol equality

18,496 sets of identical genes were identified by gene symbol equality in total.

Identical genes established by genomic coordinate agreement

The distribution of BSCCs according to gene symbol equality and genomic coordinate similarity. Among the 22,274 BSCCs discovered from Human DB from three gene databases, 16966 (76.2%) BSCCs have both bidirectional edges between EntrezGene and Ensembl and between EntrezGene and UniGene. Only 223 (1.0%) BSCCs have bidirectional EntrezGene–Ensembl edges and 5085 (22.8%) BSCCs have bidirectional EntrezGene–UniGene edges. Gene symbols are in agreement within BSCCs in 21916 (98.4%) BSCCs but not in 348 (1.6%) BSCCs. When BSCC's gene symbols are equal, 84.4% (18494/21916) showed high level of GCA (>0.9) but they are not equal, genomic coordinates are in agreement ($GCA > 0.9$) only in 63.8% (222/348) of BSCCs.

In the present study, 18494 BSCCs with gene symbol equality and high level of GCA (>0.9) are classified as 'identical' genes, which are 83.0% of all BSCCs that we discovered. As shown in Fig. 3, most BSCC showed GCA greater than 0.9. UniGene, CGB8 in Ensembl has

cross-reference to CGB5 in EntrezGene and CGB6 in Ensembl to CGB in EntrezGene.

Gene symbol agreement

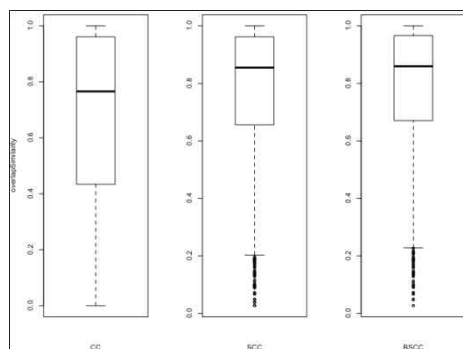


Figure 3 (upper) homogeneity at least two BSCC with CC. Y-axis means the homogeneity. (lower) the overlapSimilarity at least two BSCC with CC.

To evaluate the degree of identicalness, we measured the homogeneity of gene symbols within CC, SCC and BSCC for the whole ICN.

For the comparison between groups of CC, SCC, and BSCC, mean of homogeneity for gene symbol was calculated. As shown in Fig 2 (upper), mean of homogeneity was increased from CC, SCC, and BSCC.

Entrez Gene ID	Enterz Gene Symbol	Ensembl ID	Ensembl Gene Symbol	EntrezGene geomic coordinates	Ensembl Genomic coordinates
339010	POTEB	ENSG00000183206	POTEB, POTEC	Chr15(-): 21040700-220 83137	Chr18(-): 14511736-145 43599
378108	TRIM74	ENSG00000155428	TRIM74	Chr07(+): 75024902-750 34888	Chr07(-): 72430015-724 39997
641522	ARL17	ENSG00000185829	ARL17P1, ARL17	Chr17(-): 44376913- 44439134	Chr17(-): 44594068-446 57088
643752	hGC_1757335	ENSG00000176276	AC113404.3	Chr12(+): 69004695-690 54250	Chr05(-): 75465910-754 70171

Table 1. Matched pairs whose gene symbols are equal but genomic coordinates are completely different.

Perfect match of genomic coordinates between two entities implies high level of identicalness. Higher level of GCA in BSCC compared to CC and SCC shown in Fig. 2 indicates more probability of finding identical genes in BSCC than that in CC or SCC.

TRIM74 in EntrezGene (ID: 378108) and Ensembl (ID: ENSG00000155428) has completely different genomic coordinates

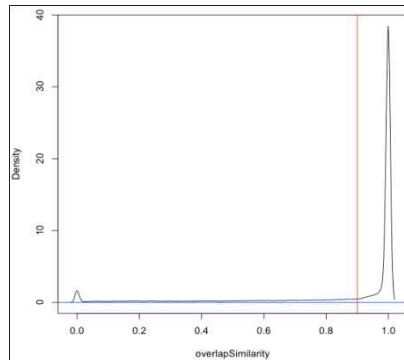


Figure 4. Density plot of genomic coordinate agreement within BSCC

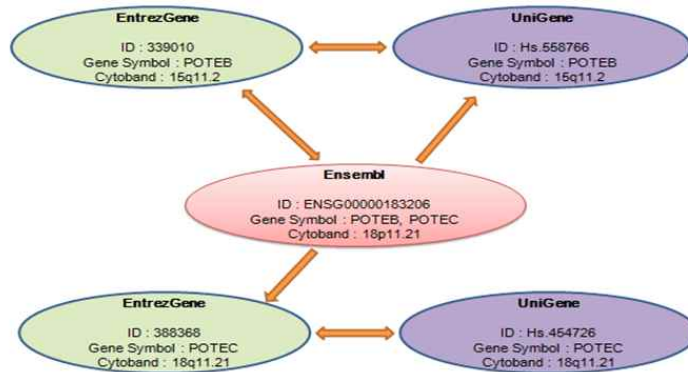


Figure 5. shows cross-reference topology around POTE B and POTE C. Our definition of BSCC correctly identified one bidirectional path among UniGene (ID: Hs.558766), EntrezGene (ID:339010), and Ensembl (ID: ENSG00000183206) entities and another between EntrezGene (ID:388368) and UniGene (ID: Hs.454726) entities. As a result, the CC in Fig. X is decomposed into two BSCCs. Notice that the Ensembl entity has two gene symbols, POTE B and POTE C, and cross-reference to EntrezGene's POTE C (ID: 388368) as well as cross-references to the two entities within the BSCC.

Identicalness

There were consistencies gene symbol consistencies and between RGGN and overlap similarity > 0.9 . These results indicate these three methods can complement each other in identifying identical genes. false positive cases may be filtered.

The danger of adopting intersection of three methods is to produce many false negative cases.

In this way, we found 18,494 sets of identical genes across three gene-centric databases (Table 1).

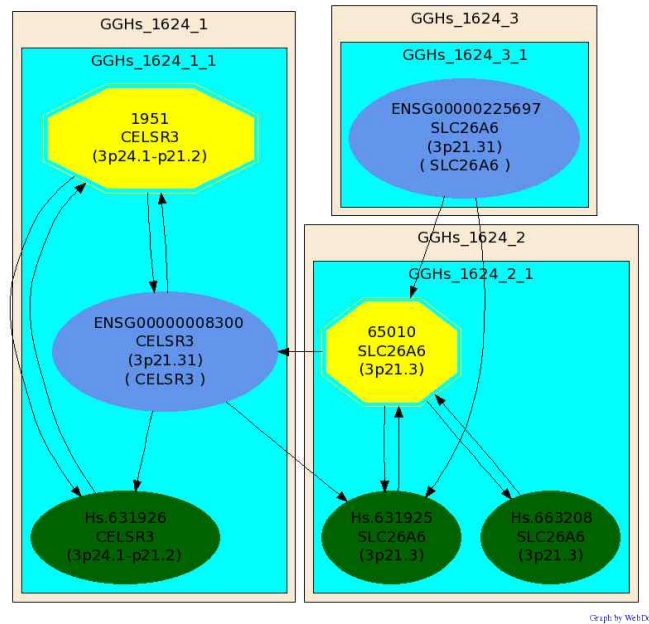


Figure 6. BSCC Examples

Among BSCCs showing gene symbol equality, five showed completely non-overlapping genomic coordinates between EntrezGene and Ensembl. There are types with reverse strands and completely different chromosomes (Table 2). This implies that the present method can be applied to find the wrong gene **symbol** annotations.

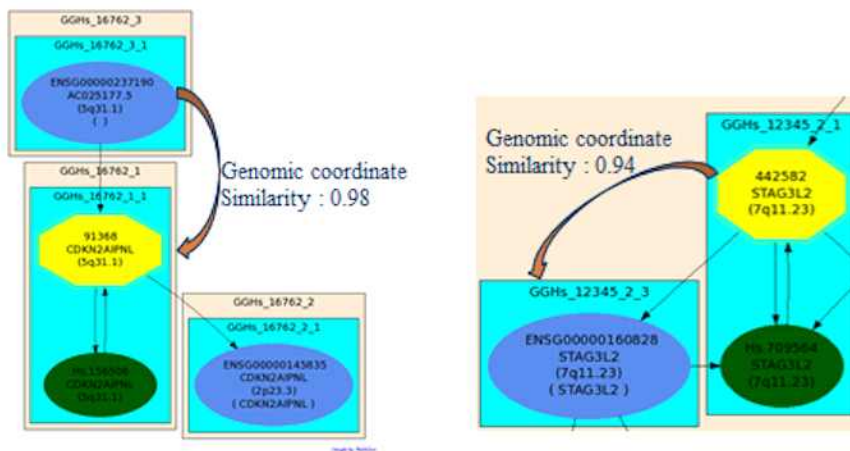


Figure 7. (a) error case, (b) merge by finding secondary BSCC after condensing primary BSCCs.

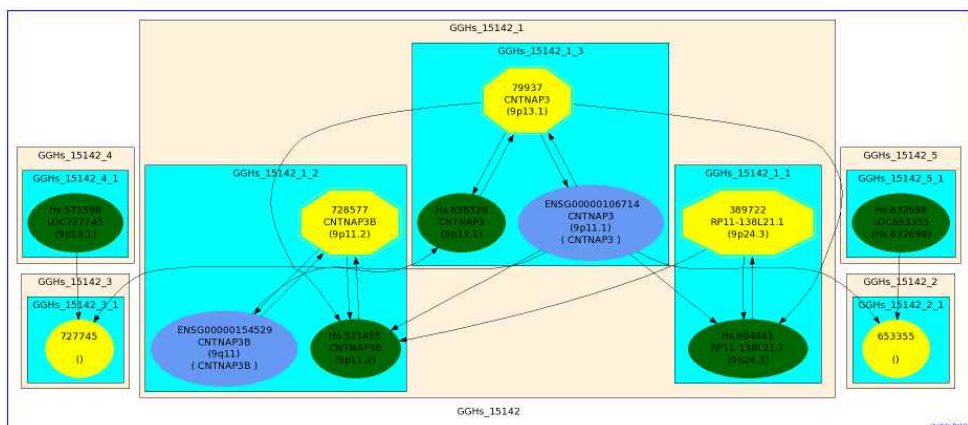


Figure 8. CC, SCC and BSCC example. Blue box indicates Connected Component. Antique white box indicates Strongly Connected Components. And cyan box indicates BSCC. There are one CC, five SCC and seven BSCC(four singleton in the case). Yellow node is EntrezGene, green node is UniGene and blue nodes is Ensembl Gene. Each node has its identifier, gene symbol and cyto-genetic location if available. The prefix GGHs_<number> is the database identifier in this system.

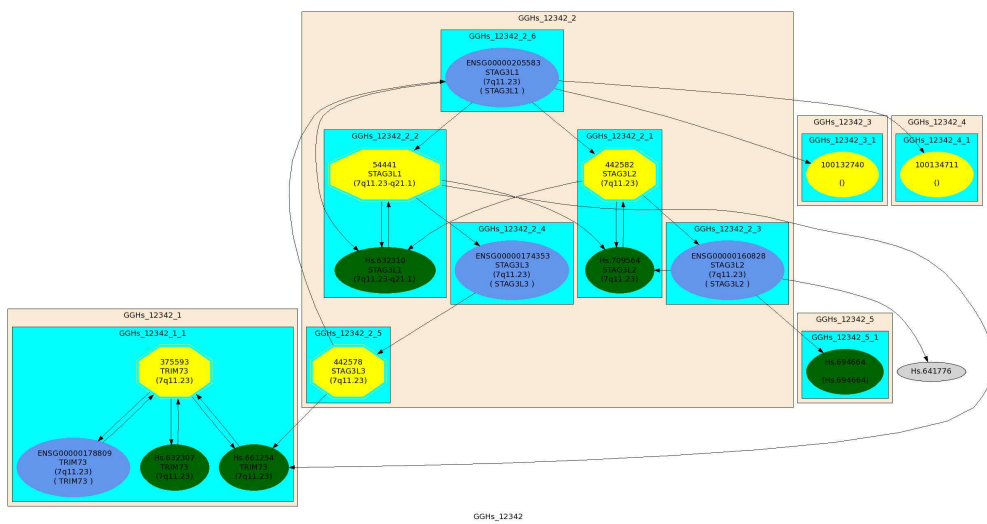


Figure 9. Singleton nodes need to be merged into another high similar BSCCs. For example ENSG00000160828(GGHs_12342_2_3) can merge into GGHs_12342_2_1 because of these have equal symbol and high genomic coordinate similarity (> 0.9). ENSG00000205583(GGHs_12342_2_6) can merge into GGHs_12342_2_2 and ENSG00000174353(GGHs_12342_2_4) can merge into GGHs_12342_2_5 also. Grey nodes(Hs.641776) is retired UniGene identifier.

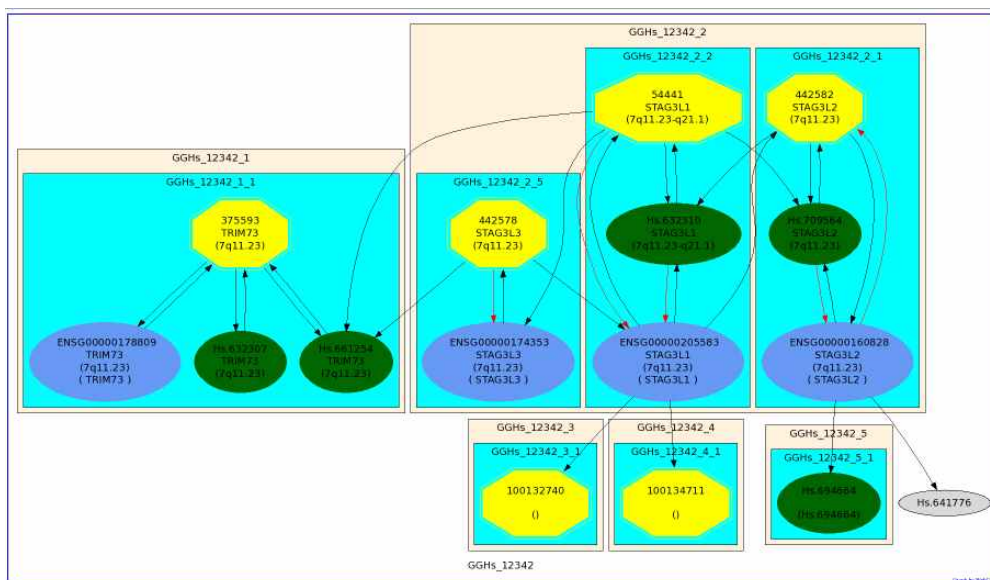


Figure 10. After merging singleton nodes. Red arrows are reconstructed

link. For example EntrezGene:442578 do not have link
Ensembl:ENSG00000174353. Our system find and curate link information.

Guided navigation across biological databases with cross reference visualization

Searching biological resources for the information of interest is an indispensable routine in current genomics and proteomics research for better understanding of experimental results and further hypothesis generation. Unfortunately, however, it is often difficult to find all associated information within a single resource, thus researchers have to search a multitude of related resources for further investigation of biological entities. Cross-references, the link information to other database identifiers representing the same biological entities or sharing common biological attributes, have been the most successful approach for the integration of biological databases (Stein, 2003) and provide valuable clues in searching for the relevant information across resources.

Identifying the counterpart IDs in a target databases from an input ID is a crucial first step in associating diverse types of data from various sources for the comprehensive explanation of complex biological systems such as interpreting gene expression profiles in the context of biological pathways (Chung *et al.*, 2004). GeneLynx, the meta-database with an extensive collection of hyperlinks to gene-specific information in diverse databases, provides a categorized listing of cross-references pertinent to a gene in tabular format (Lenhard *et al.*, 2001). This tabular representation, however, does not show how the listed IDs are interconnected, which is even more important because not all identifiers are related in the same way and to the same degree. For example, some sequence IDs are linked to a

gene through the membership of the corresponding UniGene cluster while others through that of LocusLink.

PDBSprotEC maps Enzyme Classification numbers and Protein Data Bank (PDB) using cross-references to SwissProt (Martin, 2004). GeneHopper links IDs in different expression resources based on UniGene clusters (Svensson *et al.*, 2003). These applications, however, have limitations in that their search strategies are restricted to a predefined set of cross-references and input and target IDs.

BioGPS handles cross-references using a graph structure with biological IDs as nodes and cross-references as edges. Instead of providing a rigid tabular listing of related IDs, BioGPS visualizes cross-reference graphs across diverse databases. One can interactively choose the set of databases to be searched. One can interactively pipeline the search steps, too. For example, one can easily create an emulator of PDBSprotEC by setting PDB and EC as input and target databases and SwissProt as the intermediary search path.

PROGRAM OVERVIEW

BioGPS currently serves three species: human, mouse and rat. A dozen of databases are integrated to obtain IDs and cross-references: UniGene, LocusLink, RefSeq, OMIM, Ensembl, Genew, SwissProt, TrEMBL, PROSITE, InterPro, Pfam, PDB and NetAffx. Flat files from above databases are parsed by a Python code. Each parsed identifier is treated as a node and each cross-reference a directed edge from the node holding the cross-reference to the node targeted by the cross-reference.

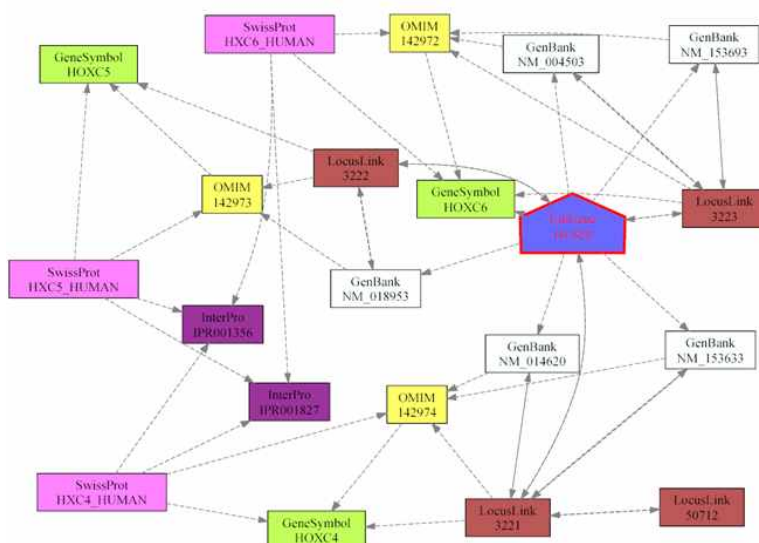


Figure 11. Outcome screenshots of ID profiling and ID mapping. Nodes are color-coded according to their source database. ID profiling of UniGene, Hs.820, LocusLink, SwissProt, OMIM, InterPro, RefSeq IDs and official gene symbols that are within three depths of cross-references, are displayed

BioGPS has two functions: ID profiling and ID mapping. First, the ID profiling function exhaustively searches the cross-reference network from a given ID for directly and indirectly linked ones. Then it visualizes the search results as a network diagram to provide an intuitive overview of the cross-reference network. The network diagram around a UniGene, Hs.820 is exemplified in Fig.1A. It is shown that Hs.820, which is labeled as 'Home boxC6' in UniGene, comprises three genes (*i.e.* HOXC4, HOXC5, HOXC6), each of which has distinct Entrez Gene ID (*i.e.* 3221, 3222 and 3223, respectively), and that their corresponding proteins (*i.e.* HXC4_HUMAN, HXC5_HUMAN and HXC6_HUMAN) contain common homeobox domains (*i.e.* IPR001356, IPR001827). Further survey informs that these genes are one of HOXC genes that are co-transcribed in a primary

transcript and are processed to gene-specific transcripts and that mRNA RefSeq IDs assigned to the same Entrez Gene ID are transcript variants of the corresponding gene.

Second, the ID mapping function resolves the counterpart ID in user-specified target database from an input ID by searching through a user-specified set of intermediary databases. The result is provided as a tree diagram with the input ID as root, the counterpart IDs as leaves and the shortest paths to the targets as stems. In addition to the resulting counterpart IDs, the tree diagram shows user how they are identified. For the purpose of illustration, PDBSprotEC is emulated by the mapping between EC number and PDB IDs using SwissProt as the intermediary path (Fig.1B). One can flexibly design useful query strategies by setting the input, intermediary and target parameters. Fig.1C shows the result of mapping all OMIM disease information associated to a biological pathway (human TCA cycle) by searching the OMIM counterpart IDs from all EC numbers found in the pathway.

GRIP(Genome Research Information Pipeline)

Searching biological databases is the essential routine that interpret the results of biological experiments and form a new hypothesis in genomics and proteomics fields.

It is difficult to retrieve all related information despite most researchers build in-houses or local databases to share information within groups. One must retrieve numerous resources to collect biological entries.(Stein 2003; Stefan 2005)

It is most challenging bottleneck that integrate enormous biological resources. in bioinformatics/biomedical research field. (Davidson 1995; Stein, 2003)

Link integration has been by far the most successful case in biological database integration (Stein 2003). It describes inter-connectivity only between source and destination database without relational database modeling. Also there is no need for to know for detailed schema of destination database. Considering source database is a record of linked set, these approaches provide efficient navigation through diverse databases(Davison 1995; Stein 2003; Stefan 2005; Hernandez and Kambhampati 2004), so major biological databases such as NCBI, DB-GET(Fujibuchi 1998) use linked record or semi-structured model. GeneLynx(Lenhard 2001) and GeneHopper(Svensson 2003) are stored only cross-reference information without biological object modeling.

But link integration has limited advantages due to 1) can not maintain integrity between source and destination database. Consider If source database has entry link into destination database entry, but

destination databases' entry was removed and vice versa, namely 'link withdrawn' 2) can not handle entry ambiguity, for example, naming crash, synonymous. 3) Each link has its biological meaning. For example, HGNC:5962 is the human interleukin 10 Gene(IL10) from HGNG(HUGO Gene Nomenclature Committee) and MGI:96537 is the mouse interleukin 10 Gene from MGI(Mouse Genome Database). The link from HGNC:5962 to MGI:96537 indicates that there are homologous relationship between two entries. So researcher has responsible to interpret its entry-entry relationship.

Data Warehousing and mediator-based integration query translation vs. data translation.

Object Oriented Modeling provides a structured data model and query environment for effective biological entity compared to the relational database modeling. In the object-oriented modeling perspective, HOWDY (Hirakawa 2002) is largely divided by the Database Object (DBO) and the Biological Objects (BO) (eg, Gene Class, SNP Class and Protein Class). DBO contains a cross-link property sharing common attributes that were extracted from 14 public databases (eg, name, title and alias). For example, gene name is extracted from the property of HUGO, Locus Link, and GDB.

These object-oriented modeling approaches make it possible to search for keywords, multiple targets for a specific object, and combination of the cytogenetic position. The GeneCard extract an integrated feature extraction from the sources such as SWISSPROT, OMIM, Gene Atlas, and GDB and define a gene using the approved gene symbol set by the HUGO / GDB nomenclature committees.

While SOURCE use the gene names and gene products defined by the UniGene cluster. Gene Keydb are based on data mining

environment provided by the Entrez Gene.

It is very difficult to manage and integrate the biology DB Warehouse because of the challenges for generating global schema for the complex biological objects and frequent changing of attributes due to the nature of biological database.

There are the same challenges for the Mediator-based integration such as integration & maintenance problem, because of the slow performance and limited sources of data.

Because Open Integration removes the relational model, there is the advantage of management and integration such as “quick-and-dirty” and “easy-to-implement”, though it does not guarantee data integrity and disambiguation. It is easy to update simply using a cross reference information and add a new database or cross references through the design of a parser.

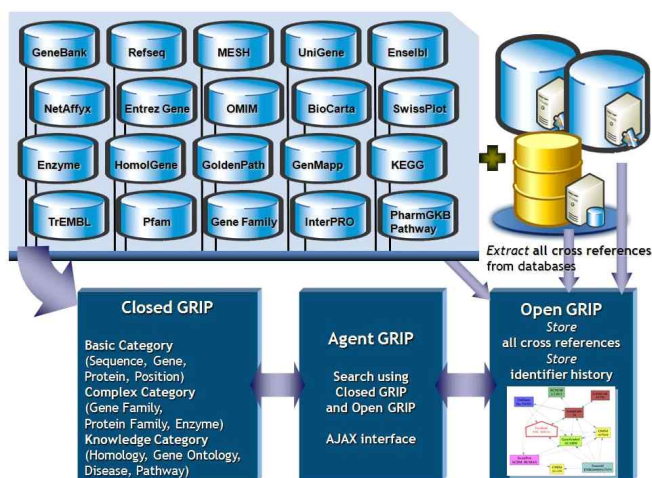


Fig 12. Overview of GRIP System

It is almost impossible to warehouse or very difficult to maintain of the whole biological databases, while the OOP Based Closed

integration provides a structured data model and query environment for the effective are biological entity. GRIP combined the benefits of open-closed-integration including OOP modeling as a balanced fashion [Figure 1].

Open GRIP created a single relational table laid for a vast network of biological entry. Open GRIP is scalable and easy to maintain for the various source.

The modeling of the Closed GRIP was based on the object of OOP modeling (eg. Sequence, gene, protein, gene family, protein family, enzyme, disease, pathway, homology) and built an warehouse by integrating the major biological databases (e.g., GenBank(Benson 2007), RefSeq(Pruitt 2005), UCSC GoldenPath(Karolchik 2003), dbSNP(Sherry 2001), NetAffx, Entrez Gene(Maglott 2005), UniGene(Pontius 2003), SwissProt/TrEMBL(Boeckmann *et al.*, 2003), HGNC gene family, Pfam(Finn 2006), OMIM, MeSH, Homologene, KEGG(Kanehisa 2004), BioCarta(<http://www.biocarta.com>) and GenMAPP(Dahlquist 2002)).

Closed GRIP provides a unified data model and query efficient environment that includes all the information contained in the major biological database.

Agent GRIP, collection of script programs, makes the Open and Closed GRIP work together. Therefore GRIP makes extensive search for a biological object of any level, such as DNA Microarray studies.

We have successfully implemented and tested the GRIP Produced in the same system that were tested for the previous tools such as ArrayXPath I (Chung 2004) and II (Chung 2005), ChromoViz (Kim 2004), GOChase (Yu Rang Park 2005) and Xperanto (Ji Yeon Park 2005).

METHODS

GRIP modeled ten biological objects divided by three categories, basic, complex, and knowledge. The ten biological objects were sequence,

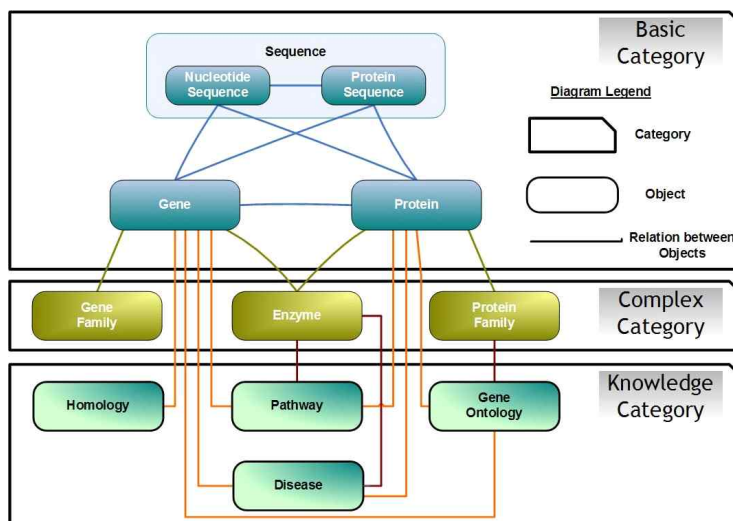


Fig 13. The Categories of GRIP Objects

gene, protein, gene family, protein family, enzyme, disease, homology, Gene Ontology, and pathway. [Figure 13]

Each Object were saved as the EAV (Entity-Attribute-Value) format and the relationships of the Object were defined by extracted attributes.

Table 1 showed the status for each object, Source DB, and extracted attributes.

Object Category

There were three types of categories, basic, complex and knowledge categories according to the characteristic of each Object.

Basic Category

Basic category composed of four basic objects having a physical nature such as sequence, gene, protein and so on.

Complex Category

Complex category is composed with gene-family, protein-family, and enzyme objects that can be represented as a set of basic category.

Gene family object was constructed using gene group (<http://www.gene.ucl.ac.uk/nomenclature/genefamily.html>) established in HGNC for human and using gene group (<http://www.informatics.jax.org/mgihome/nomen/genefamilies/index.shtml>) established in MGI for mouse.

Protein family object was constructed by reference to the Pfam and InterPro containing structural and functional information related to the protein.

Enzyme object consists of a set that is composed of the subsets from Entrez Gene and SwissProt ENZYME of the gene / protein object provided by KEGG metabolic pathway.

Knowledge Category

Knowledge category was consisted of a set of basic and/or complex category object. The knowledge category objects were modeled as a Knowledge-level, such as biological pathway, homology, and disease.

The pathway object integrated gene, protein, enzyme object and pathways including KEGG (the 417 for human 471, 277 for mouse 277, 64 for rat), GenMAPP, BioCarta, and PharmGKB pathway (28).

And Homo sapiens are consists of 3,151 genes (or gene products) and 121 enzyme. We referenced the HomoloGene for the Homology object, the Entrez Gene, SwissProt, and InterPro for the Gene Ontology Object to extract entities.

Disease object contains a disease-associated gene and/or protein objects.

There are C category in the MeSH (Medical Subject Headings) which contains 23 branches and have entry terms and disease names as a hierarchical structure.

MESH and the OMIM Morbid Map is extracted gene-related/disease-related term from the (<http://www.ncbi.nlm.nih.gov/Omim/getmorbidity.cgi>).

Disease names were mapped by the exact keyword match method through the MESH. There were 3,259 official gene symbols resolved from the Morbid Map. Among these genes, 2,395 genes were mapped through MeSH heading or entry term. Disease-related protein was extracted through the Disease term of the Comment Entry of SwissProt.

Category	Objects	Source DB	Extracted attributes
Basic	Sequences	GenBank, RefSeq, NetAffx, dbSNP	Organism, Definition, Accession ID, GI, AffyMetrix Probe ID, rs id
	Gene	EntrezGene	Entrez Gene ID, Gene Name, Official Gene Symbol, Alias Gene Symbol, Chromosome, Cytoband, locus tag, region, SUMMARY, UniGene ID
		UniGene	UniGene ID, Gene Name, Gene Symbol, Cytoband, Clone ID, Locus ID
		Ensembl	Gene Name, Gene ID, Transcript ID, Peptide ID, Exon ID, Entrez Gene ID, UniGene ID
	Protein	SwissProt, TrEMBL	SwissProt ID, primary accession no, DE, GN, CC, DR
	Physical Position	UCSC Golden Path	ID(sequence, gene), chromosome, strand, begin, end
Complex	Gene Family	HGNC Gene Family	Title, Root Symbol, HGNC Link, Link Name, Status
	Protein Family	Pfam	Pfam ID, Pfam acc no., Type, Description, DR
	Enzyme	SwissProt-Enzyme	AN, CA, CC, CF, DI,, DR, PR
Knowledge	Homology	HomoloGene	HomoGene ID, Gene Symbol
	Gene Ontology	Entrez Gene, SwissProt, InterPro	GO ID, Evidence, Description
	Disease	OMIM, MeSH	OMIM ID, Title, Disease Name, Entry Term
	Biological Pathway	BioCarta, GenMapp, KEGG, PharmGKB	Pathway Name, Gene ID

Table 2. Categories, Class, Objects and Attirbutes in GRIP System

THE IMPORTANCE OF GENE OBJECT & INTEGRATE MULTIPLE GENE RELATED DB

Gene Object is the hub between all close GRIP Objects

Gene object has an important role for the closed integration of object-oriented modeling and biological object of GRIP, that means as a hub.

Gene object is connected to the protein object and the sequence object and thereby helps to connect other complex / knowledge category objects in the GRIP.

GRIP generated the unique gene object systematically using cross-referenced information between the references of Entrez Gene and UniGene unlike GeneCards, GeneKeyDB or SOURCE which depends only one of them.

OPEN INTEGRATION BASED ON CROSS REFERENCE

Structure of Open Integration

Open GRIP to create a network consisting of a biological database identifier associated with the cross-reference. Network were configured as a single table which reflects five columns ([Start DB name] - [Start DB ID] - [End DB name] - [End DB ID] - [Source name]) reflects a directed graph. For example, when Hs.2 in UniGene flat file refers to GenBank NM_000015, add a row as [UniGene] - [Hs.2] - [GenBank] - [NM_000015] - [UniGene].

Identifier History Tracing

Unique Identifier played a major role as the 'Key Abstraction' or 'scatterfold' served to determine the relationship between the different databases and Biological Entity (Andrew 2007)

However, the identifier could be replaced or discarded when a database updates. If the microarray created based on the UniGene Cluster has only UniGene Cluster ID without the version information, it would be difficult to trace the corresponding sequence after update of UniGene DB.³⁴

In the Open GRIP, the identifier stored whenever the databases are updated. The GRIP would be a very useful tool to trace and track the identifiers though the incremental storage of changes of them.

AGENT GRIP

The most important characteristic of Agent GRIP is a function that allows searching through the two GRIP.

Agent GRIP performed the following procedures when receiving the user queries:

1. Search using Closed GRIP
2. Search using Open GRIP
3. Display the query results

Searching using Closed GRIP search target

GRIP
Genome Research Information Pipeline

Home | Object Diagram | **Keyword Search** | Advanced Search | Statistics | About

Menu
Home
Object Diagram
Keyword Search
Advanced Search

Statistics
Sequence
Gene/Gene Family
Protein/Protein Family
GO/Pathway/Disease

What is GRIP?
GRIP is a web-based integrated database for analyzing and accessing biological information of human, mouse and rat

Last Update
29 June 2007

KCNQ1
Input : KCNQ1

Gene (1 found)

Symbol	KCNQ1
Entrez Gene:	3784 UniGene: Hs.95162 Ensembl: ENSG00000053918 ENSM00000155840 ENSM00000380776 ENSM00000335475 ENSM00000345015 ENSP00000155840 ENSP00000370153 ENSP00000334497 ENSP00000342896
External Link	
Name	potassium voltage-gated channel, KQT-like subfamily, member 1 (from Entrez Gene) Potassium voltage-gated channel, KQT-like subfamily, member 1 (from UniGene) Potassium voltage-gated channel subfamily KQT member 1 (Voltage-gated potassium channel subunit Kv7.1) (Ks producing slow voltage-gated potassium channel subunit alpha KvLQT1) (KQT-like 1) [Source: Uniprot/SwissProt; Acc:P51787] (from Ensembl)

Figure 14 showed that the 'KCNQ1' as input the results found in the results (left) Closed GRIP and the Agent GRIP Display by Object. The figure in the upper left corner of Figure 3 represents the nodes and edges on the basis of the search results of the Closed GRIP. Picture on the top right showed the id profiling network configured by the Open GRIP (Figure 3).

When users input the 1LJD as of PDB id and click the found results (right) node in the Agent GRIP will be moved back to the information of the Closed GRIP. If there is no available information in the Closed GRIP, the browser moves on to the original site for the ID. When a

searches exhaustively the biological object C-GRIP. All information of the biological objects having association with retrieved biological was represented in tables and graphs.

And the id-profiling information of input key was visualized through the O-GRIP by graphs.

For example, a search for 'KCNQ1', the description of the gene was represented in C-GRIP by GRIP. And also the information associated with that sequence, protein, pathway, disease, homology and pathway related to the 'KCNQ1' were expressed as tables. In addition, the protein family information of the gene, 'KCNQ1' was displayed in tables for each object associated with the protein. Finally, visualize the associated information as a navigation graph of biological objects.

The cross-reference relationships also displayed by querying the O-GRIP relation to the 'KCNQ1' with the id-profiling graph [Figure 3 (left)].

Searching using Open GRIP search target

The next scenario is that the input id is not found in integrated databases in the C-GRIP or previous version id.

In this case, GRIP provides an id profiling through the O-GRIP.

The id matching process is 1) The users use the id-profiling graph of the O-GRIP, 2) explore graph-based navigation, and 3) select for the C-GRIP ID. The first scenario started after this process.

BIOLOGICAL KNOWLEDGE-BASED SEARCH

Biological knowledge-based search is a process of searching using biological attributes on more than one biological object and annotation of the results of biological experiments.

Users can define the desired search by combining each object in GRIP. For example, the process to search for a list of genes related to lipid metabolism in the human chromosome 11 is as follows:

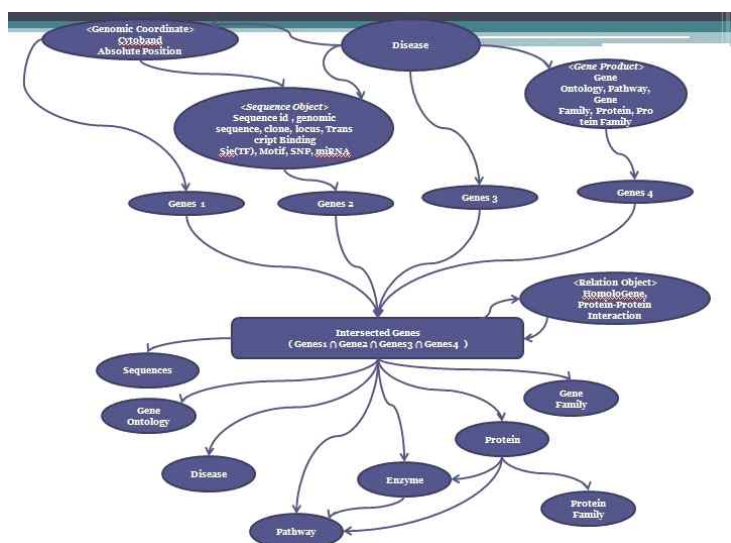


Figure 16. Knowledge Based Search Scheme.

First, select No. 1 chromosome from the 'Physical Position' Object. Then select a 'lipid metabolism' in the Gene Ontology Object. GRIP will show the gene ontology.

Select 'lipid metabolism (GO: 0006629)' and then 'drag & drop' the Position object and Gene Ontology Object in "customize your search" box.

Finally, a check for 'Gene' in the Output and click the 'search' button to be able to see the results of the gene set in that condition.

DISCUSSION

Identifying identity between genes is a key problem in integration of gene-centric databases. Decision on the identity of genes is usually considered as what domain experts have to do. But it is tedious, time-consuming, and hard to achieve a consensus even between domain experts. In this context, computational methods to identify identical genes are demanding. Traditionally gene symbol and genomic location information could be reliably used for decision on gene identity problem. However as shown in the above example, they are often unreliable.

Introduction of cross reference information were used in identifying identical genes in some previous researches. However in all of these researches, reliability of the cross reference information was not considered. Our study is the first one where the network topology was used to increase the reliability of cross reference information.

RGGN is produced by compromising the different perspectives on genes the three gene-centric databases have. The probability would be very low that the falsely inserted cross reference information produces many new BSCCs.

Gene was not factitiously defined, but data about genes were naturally considered. This scheme can be universally used for any entities with multiple compelling perspectives.

In the actual implementation of these schemes, to perform the process of construction of RGGN before filtering by overlap similarity > 0.9 is computationally advantageous because calculation of overlap similarity is very extensive.

Our data could be used for correcting errors of cross referencing in each of EntrezGene, UniGene, and Ensembl. Through these processes, these databases can provide more reliable data to the user and the quality of data in our database will also be increased.

These schemes can be also used for actual integration of gene-centric databases. Because of high reliability of this scheme, it can be used as closed integration scheme. For the rest of genes not included in these schemes, we can use open integration methods. Use of two integration methods can produce both high flexibility and high reliability for the query results. we made Genome Resource Annotation Pipeline (GRIP), where various entities ranging from genes, proteins, diseases are interactively annotated in gene-centric ways. GRIP has three main parts, closed GRIP, open GRIP, and agent GRIP, among which closed GRIP was made based on these schemes.

Because our schema is conceptually concrete and computationally feasible, we can update our data following update schedule of any of EntrezGene, UniGene, or Ensembl.

REFERENCES

Andrew K Smith, Kei-Hoi Cheung, Kevin Y Yip, Martin Schultz and Mark B Gerstein LinkHub: a Semantic Web system that facilitates cross-database queries and information retrieval in proteomics BMC Bioinformatics 2007, 8(Suppl 3):S5

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2007 Jan;35(Database issue):D21-5.

Brigham H. Mecham, et.al . Increased measurement accuracy for sequence-verified microarray probes. Physiol Genomics 18:308-315, 2004

Brigitte Boeckmann et. al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research, 2003, Vol. 31, No. 1 365-370

Chung,H.J., Kim,M., Park,C.H., Kim,J. and Kim,J.H. (2004) ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics, *Nucleic Acids Res.*, **32**, W460-446.

Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nature Genet., 31, 19 - 20.

Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova

Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res. 2005 January 1; 33(Database Issue): D54 - D58.

Dov Stekel , Microarray Bioinformatics, Cambridge University Press, UK; 2003; ISBN 0-521-52587-X; 33 pp

Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa, M. (1998) DBGET/LinkDB: an integrated database retrieval system, Pac. Symp. Biocomput., 683-694

Hee-Joon Chung, Chan Hee Park, Mi Ryung Han, Seokho Lee, Jung Hun Ohn, Jihoon Kim, Jihun Kim and Ju Han Kim. (2005) ArrayXPathII: mapping and visualizing microarray gene-expression data with biomedical ontologies and in-tegrated biological pathway resources using Scalable Vector Graphics, Nucleic Acids Res., 33, W621-W626

Hee-Joon Chung, Mingoo Kim, Chan Hee Park, Jihoon Kim and Ju Han Kim. (2004) ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using Scalable Vector Graphics, Nucleic Acids Res., 32, W460-W464

H. Pearson(2006) Genetics: What is a gene? Nature 441, 398-401(2006)

Mark B. Gerstein, Can Bruce, Joel S. Rozowsky, Deyou Zheng, Jiang Du, Jan O. Korbelt, Olof Emanuelsson, Zhengdong D. Zhang, Sherman Weissman, and Michael Snyder(2007) What is a gene, post-ENCODE? History and updated definition, Genome Res. 2007. 17: 669-681

Ji Yeon Park, Yu Rang Park, Chan Hee Park, Ji Hoon Kim and Ju Han Kim. (2005) Xperanto: a web-based integrated system for DNA microarray data management and analysis, *Genomics and Informatics*, 3(1), 39-42

Jihoon Kim, Hee-Joon Chung, Chan Hee Park and Ju Han Kim. (2004) Chro-moViz: mutlimodal visualization of gene expression profile onto chromosome using scalable vector graphics, *Bioinformatics*, 20, 1991-1992

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*,32, D277 - D280.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. and Kent, W.J. The UCSC Genome Browser Database. *Nucl. Acids Res* 31(1), 51-54 (2003).

Kohler, J., Philippi, S. and Lange, M. (2003) SEMEDA: ontology based semantic integration of biological databases, *Bioinformatics*, **19**, 2420-2427.

Lenhard,B., Hayes,W.S. and Wasserman,W.W. (2001) GeneLynx: a gene-centric portal to the human genome, *Genome Res.*, **11**, 2151-2157.

Mark B. Gerstein, Can Bruce, Joel S. Rozowsky, Deyou Zheng, Jiang Du, Jan O. Korbel, Olof Emanuelsson, Zhengdong D. Zhang, Sherman Weissman, and Michael Snyder. What is a gene, post-ENCODE? History and updated definition

Martin,A.C. (2004) PDBSprotEC: a Web-accessible database linking PDB chains to EC numbers via SwissProt, *Bioinformatics*, **20**, 986-988.

Maximilian Diehn, Gavin Sherlock, Gail Binkley, Heng Jin, John C. Matese, Tina Hernandez-Boussard, Christain A. Rees, J. Michael Cherry, David Botstein, Pat-rick O. Brown and Ash A. Alizade h. (2003) SOURCE: a unified genomic re-source of functional annotations, ontologies, and gene expression data, *Nucleic Acids Res.*, 31, 219-223

McKusick, V.A.: Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. Baltimore: Johns Hopkins University Press, 1998 (12th edi-tion).

Michael rebhan, Vered Chalifa-Caspi, Jaime Prilusky and Doron Lancet. (1998) GeneCards: a novel functional genomics compendium with automated data min-ing and query reformulation support, *Bioinformatics*, 14(8), 656-664

Mika Hirakawa. (2002) HOWDY: an integrated database system for human genome research, *Nucleic Acids Res.*, 30, 152-157

NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins Pruitt KD, Tatusova, T, Maglott DR *Nucleic Acids Res* 2005 Jan 1;33(1):D501–D504

Pfam: clans, web tools and services Robert D. Finn, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R. Eddy, Erik L. L. Sonnhammer and Alex Bateman *Nucleic Acids Research* (2006) Database Issue 34:D247–D251

Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcrip-tome. In: *The NCBI Handbook*. Bethesda (MD): National Center for Biotechnol-ogy Information; 2003.

Rudi Alberts et. al. A verification protocol for the probe sequences of Affymetrix genome arrays reveals high probe accuracy for studies in mouse, human and rat *BMC Bioinformatics* 2007, 8:132

S.B.Davidson, C.Overton and P.Buneman. (1995) Challenges in integrating biolog-ical data sources. *J Comput Biol.*, 2(4), 557–572

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B.B., Butler, A., Castle, A.B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P.J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Edwards, C., Fan, J.B., Fang, N., Fizames, C., Garrett, C., Green, L., Hadley, D., Harris, M., Harrison, P., Brady, S., Hicks, A., Holloway, E., Hui, L., Hussain, S.,

Louis-Dit-Sully, C., Ma, J., MacGilvery, A., Mader, C., Maratukulam, A., Matisse, T.C., McKusick, K.B., Morissette, J., Mungall, A., Muselet, D., Nusbaum, H.C., Page, D.C., Peck, A., Perkins, S., Piercy, M., Qin, F., Quackenbush, J., Ranby, S., Reif, T., Rozen, S., Sanders, C., She, X., Silva, J., Slonim, D.K., Soderlund, C., Sun, W.L., Tabar, P., Thangarajah, T., Vega-Czarny, N., Vollrath, D., Voyticky, S., Wilmer, T., Wu, X., Adams, M.D., Auffray, C., Walter, N.A., Brandon, R., Dehejia, A., Goodfellow, P.N., Houlgatte, R., Hudson, J.R., Jr., Ide, S.E., Iorio, K.R., Lee, W.Y., Seki, N., Nagase, T., Ishikawa, K., Nomura, N., Phillips, C., Polymeropoulos, M.H., Sandusky, M., Schmitt, K., Berry, R., Swanson, K., Torres, R., Venter, J.C., Sikela, J.M., Beckmann, J.S., Weissenbach, J., Myers, R.M., Cox, D.R., James, M.R., Bentley, D., Deloukas, P., Lander, E.S. and Hudson, T.J. (1996) A gene map of the human genome, *Science*, **274**, 540-546.

S. T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin dbSNP: the NCBI database of genetic variation *Nucleic Acids Research*, 2001, Vol. 29, No. 1 308-311

Stefan A Kirov, Xinia Peng, Erich Baker, Denise Schmoyer, Bing Zhang and Jay Snoddy. (2005) GeneKeyDB: A lightweight, gene-centric, relational database to support data mining environments, *BMC Bioinformatics*, 6:72

Stein,L.D. (2003) Integrating biological databases, *Nature Reviews Genet.*, **4**, 337-345.

Sujansky, W. (2001) Heterogeneous database integration in biomedicine, *J Biomed Inform*, **34**, 285-298.

Svensson,B.A., Kreeft,A.J., van Ommen,G.J., den Dunnen,J.T. and Boer,J.M. (2003) GeneHopper: a web-based search engine to link gene-expression platforms through GenBank accession numbers, *Genome Biol.*, **4**, R35.

T. J. P. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle and P. Flicek(2009) Ensembl 2009, *Nucl. Acids Res.* (2009) 37 (suppl 1): D690-D697.

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. and Wagner, L. (2003) Database resources of the National Center for Biotechnology, *Nucleic Acids Res*, **31**, 28-33.

논문 초록

유전체 변이의 임상적 영향 평가를 위한 통합 유전체 데이터베이스

박 찬 희

협동과정 생물정보학 전공

서울대학교 대학원

관심 있는 biological 정보를 검색한다는 것은 각종 실험결과의 해석과 더 나아가 가설자체를 새로 세우는데 있어서 인간 게놈 분석에서 없어서는 안 되는 필수불가결한 루틴이다. 연구자들은 협력자들과 함께 정보를 쉽게 공유하기 위해 자체 데이터베이스를 만들어낸다. 그럼에도 불구하고 모든 관련된 정보를 하나의 자원 안에서 찾는 것은 매우 어렵다. 결국 연관된 생물학적 정보를 찾기 위해 수많은 관련된 데이터베이스들을 검색해야만 한다.

특히 유전자에 대한 해석을 위해서는 다양한 유전자 중심의 데이터베이스를 통합해야 한다. 하지만 각기 다른 유전자 중심의 데이터베이스 통합 시 일관성 있는 통합에 어려움이 따르는데 가령 유전자 심볼에 해당하는 유전자의 위치 등의 정보들이 각 데이터베이스에 따라 상이하게 다른 경우 등을 들 수 있다.

이 논문에서는 각 데이터베이스간의 상호참조정보, 유전자의 게놈상의 위치정보, 유전자 심볼 정보를 이용하여 생물정보학 분야에서 가장 많이 사용되는 3개의 유전자 중심 데이터베이스를 일관 되게 통합할 수 있는 방법을 제시한다. 방향성을 고려하지 않은 상호참조정보를 이용하여 유전자-유전자 상호참조정보 네트워크(GGN)를 구성하고, 각 GGN 마다 방향성 정보를 고려한 ‘안정된 유전자-유전자 상호참조정보 네트워크’(RGGN)을 구성한다. 이를 ‘Closed Integration’이라 부른다.

‘Open Integration’ 통합 방법은 Closed Integration 방법과는 달리 상호 참조 정보만을 이용하여 상호 참조 네트워크를 구성하여 수많은 생물학적 데이터베이스를 통합하는 방식이다. 생물학적 오브젝트들 간의 관계 모델링 없이 상호 참조 정보만을 저장하기 때문에 많은 수의 데이터베이스를 쉽게 통합할 수 있으며 식별자 검색 및 심볼 검색 등 키워드 정보 검색 시 매우 유용하다.

위의 두 가지 통합 방식을 이용하여 생물학 정보 및 인간 게놈 유전체 정보들을 유전자 중심 관점에서 검색할 수 있는 GRIP(Genome Resource Annotation Pipeline) 이라는 시스템을 구축하였다. GRIP은 총 10가지 생물학적 오브젝트와 총 3개(basic, complex, knowledge)의 카테고리로 나눈 구조를 취하고 있다. 키워드 기반의 검색을 제공하며, 두 개 이상의 생물학적 오브젝트를 합쳐서 검색할 수 있도록 하는 knowledge 기반 검색을 제공한다.

GRIP은 마이크로 어레이 실험 결과 및, exome/rna seq, 그리고 개인 서열 정보를 유전자 중심의 관점에서 해석할 수 있도록 연구자들에게 도움을 줄 수 있다.

키워드 : biological database, database integration,
personal genome sequence

학 번 : 2006-30132